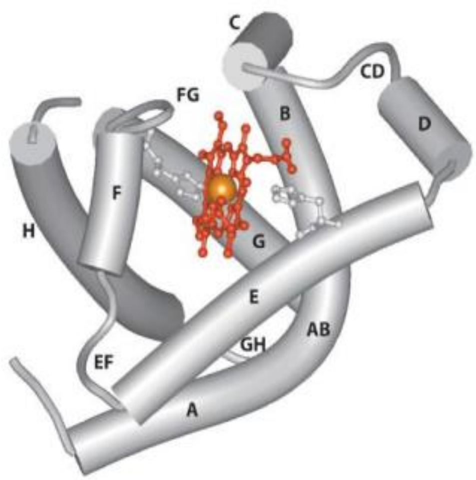
LECTURE 07: Tertiary Structure: Myoglobin and Hemoglobin

Myoglobin and hemoglobin are hemeproteins whose physiological importance is principally related to their ability to bind molecular oxygen.

Myoglobin

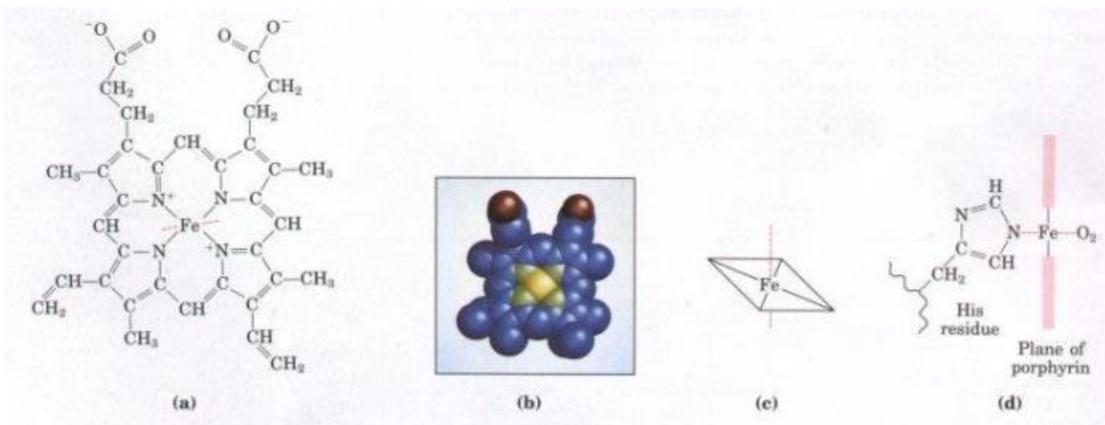
Single polypeptide chain (153 amino acids)

No disulfide bonds 8 right handed alpha helices form a hydrophobic pocket which contains heme molecule protective sheath for a heme group



Myoglobin is a monomeric heme protein found mainly in muscle tissue where it serves as an intracellular storage site for oxygen During periods of oxygen deprivation oxymyoglobin releases its bound oxygen which is then used for metabolic purposes The tertiary structure of myoglobin is that of a typical water soluble globular protein Its secondary structure is unusual in that it contains a very high proportion (75%) of α-helical secondary structure A myoglobin polypeptide is comprised of 8 separate right handed a-helices, designated A through H, that are connected by short non helical regions Amino acid R-groups packed into the interior of the molecule are predominantly hydrophobic in character while those exposed on the surface of the molecule are generally hydrophilic, thus making the molecule relatively water soluble

Each myoglobin molecule contains one heme prosthetic group inserted into a hydrophobic cleft in the protein Each heme residue contains one central coordinately bound iron atom that is normally in the Fe 2+ , or ferrous, oxidation state The oxygen carried by hemeproteins is bound directly to the ferrous iron atom of the heme prosthetic group



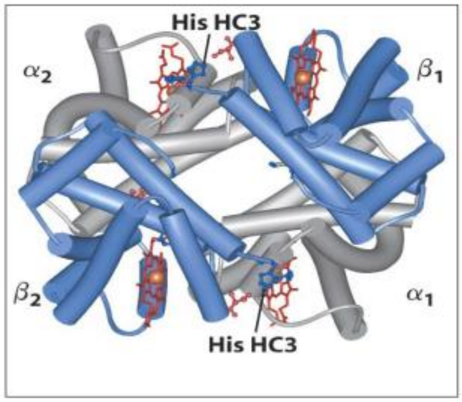
The heme group is located in a crevice Except for one edge, non polar side chains surround the heme Fe 2+ is octahedrally coordinated Fe 2+ covalently bonded to the imidazole group of histidine 93 (F8) O 2 held on the other side by histidine 64 (E7)

Hydrophobic interactions between the tetrapyrrole ring and hydrophobic amino acid R groups on the interior of the cleft in the protein strongly stabilize the heme protein conjugate. In addition a nitrogen atom from a histidine R group located above the plane of the heme ring is coordinated with the iron atom further stabilizing the interaction between the heme and the protein. In oxymyoglobin the remaining bonding site on the iron atom (the 6th coordinate position) is occupied by the oxygen, whose binding is stabilized by a second histidine residue Carbon monoxide also binds coordinately to heme iron atoms in a manner similar to that of oxygen, but the binding of carbon monoxide to heme is much stronger than that of oxygen.

The preferential binding of carbon monoxide to heme iron is largely responsible for the asphyxiation that results from carbon monoxide poisoning.

Hemoglobin

Oxygen transporter Four polypeptide chains Tetramer Each chain has a heme group Hence four O 2 can bind to each Hb Two alpha (141 amino acids) and two beta (146 amino acids) chains



Hemoglobin is an [α(2):β(2)] tetrameric hemeprotein found in erythrocytes where it is responsible for binding oxygen in the lung and transporting the bound oxygen throughout the body where it is used in aerobic metabolic pathways Each subunit of a hemoglobin tetramer has a heme prosthetic group identical to that described for myoglobin. Although the secondary and tertiary structure of various hemoglobin subunits are similar, reflecting extensive homology in amino acid composition, the variations in amino acid composition that do exist impart marked differences in hemoglobin's oxygen carrying properties In addition, the quaternary structure of hemoglobin leads to physiologically important allosteric interactions between the subunits, a property lacking in monomeric myoglobin which is otherwise very similar to the α-subunit of haemoglobin

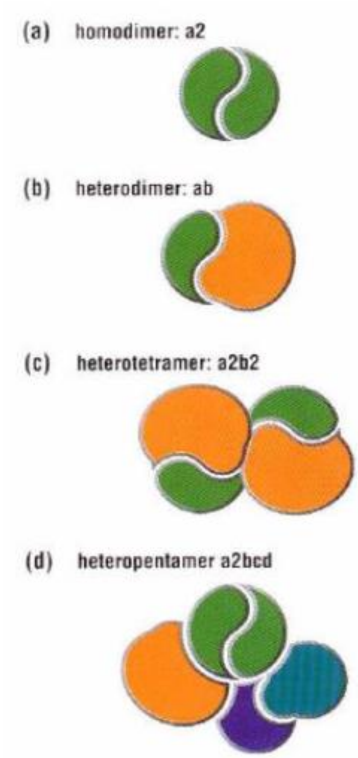
1.3. Quaternary structure

• 3-dimensional relationship of the different polypeptide chains (subunits) in a multimeric protein, the way the subunits fit together and their symmetry relationships

• only in proteins with more than one polypeptide chain; proteins with only one chain have no quaternary structure.)

Terminology

• Each polypeptide chain in a multichain protein = a subunit • 2-subunit protein = a dimer, 3 subunits = trimeric protein, 4 = tetrameric • homo(dimer or trimer etc.): identical subunits • hetero(dimer or trimer etc.): more than one kind of subunit (chains with different amino acid sequences) • different subunits designated with Greek letters – e.g., subunits of a heterodimeric protein = the "α subunit" and the "β subunit".



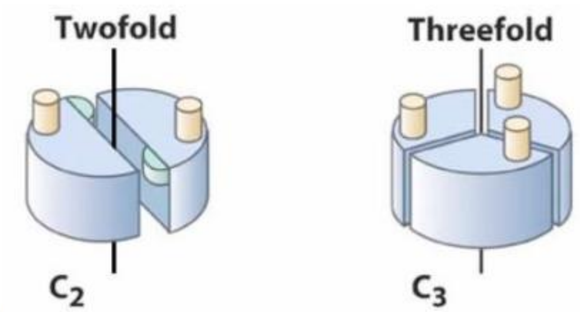
– NOTE: This use of the Greek letters to differentiate different polypeptide chains in a multimeric protein has nothing to do with the names for the secondary structures α helix and β conformation.

• Some protein structures have very complex quaternary arrangements; e.g., mitochondrial ATP synthase, viral capsids….

Symmetry in quaternary structures

• simplest kind of symmetry = rotational symmetry

• Individual subunits can be superimposed on other identical subunits (brought into coincidence) by rotation about one or more rotational axes.

.)


Two types of cyclic symmetry

• If the required rotation = 180° (360°/2), protein has a 2-fold axis of symmetry (e.g., Cro repressor protein above).

• If the rotation = 120° (360°/3), e.g., for a homotrimer, the protein has a 3-fold symmetry axis. Rotational symmetry in proteins: Cyclic symmetry: all subunits are related by rotation about a single n-fold rotation axis (C2 symmetry has a 2-fold axis, 2 identical subunits; C3 symmetry has a 3-fold axis, 3 identical subunits, etc.)

Example: Protein Capsid

Viral genomes are surrounded by protein shells known as capsids. One interesting question is how capsid proteins recognize viral, but not cellular RNA or DNA. The answer is that there is often some type of "packaging" signal (sequence) on the viral genome that is recognized by the capsid proteins. A capsid is almost always made up of repeating structural subunits that are arranged in one of two symmetrical structures, a **helix** or an **icosahedron**. In the simplest case, these "subunits" consist of a single polypeptide. In many cases, however, these **structural subunits (also called protomers)** are made up of several polypeptides. Both helical and icosahedral structures are described in more detail below.

PROTEIN ENGINEERING

1) **Helical Capsids**: The first and best studied example is the plant tobacco mosaic virus (TMV), which contains a SS RNA genome and a protein coat made up of a single, 17.5 kd protein. This protein is arranged in a helix around the viral RNA, with 3 nt of RNA fitting into a groove in each subunit. Helical capsids can also be more complex, and involve more than one protein subunit.

A helix can be defined by two parameters, its amplitude (diameter) and pitch, where pitch is defined as the distance covered by each turn of the helix. **P = m x p**, where m is the number of subunits per turn and p is the axial rise per subunit. For TMV, m = 16.3 and p= 0.14 nm, so P=2.28 nm. This structure is very stable, and can be dissociated and re-associated readily by changing ionic strength, pH, temperature, etc. The interactions that hold these molecules together are non-covalent, and involve H-bonds, salt bridges, hydrophobic interactions, and vander Waals forces.

Several families of animal virus contain helical nucleocapsids, including the *Orthomyxoviridae* (influenza), the *Paramyxoviridae* (bovine respiratory syncytial virus), and the *Rhabdoviridae* (rabies). All of these are enveloped viruses (see below).

2) **Icosahedral Capsids**: In these structures, the subunits are arranged in the form of a hollow, quasi spherical structure, with the genome within. An icosahedron is defined as being made up of **20 equilateral triangular faces** arranged around the surface of a sphere. They display **2-3-5 fold symmetry** as follows:

- an axis of 2 fold rotational symmetry through the center of each edge.

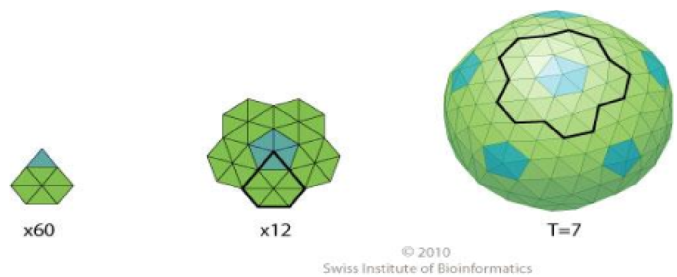
- an axis of 3 fold rotational symmetry through the center of each face.

- an axis of 5 fold rotational symmetry through the center of each corner.

These corners are also called Vertices, and each icosahedron has 12.

Since proteins are not equilateral triangles, each face of an icosahedron contains more than one protein subunit. The simplest icosahedron is made by using 3 identical subunits to form each face, so the minimum # of subunits is 60 (20 x 3). Remember, that each of these subunits could be a single protein or, more likely, a complex of several polypeptides.

Many viruses have too large a genome to be packaged inside an icosahedron made up of only 60 polypeptides (or even 60 subunits), so many are more complicated. In these cases, each of the 20 triangular faces is divided into smaller triangles; and each of these smaller triangles is defined by 3 subunits. However, the total number of subunits is always a multiple of 60. The total number of subunits can be defined as 60 X N, where N is sometimes called the **Triangulation Number**, or T. Values for T of 1,3,4,7,9, 12 and more are permitted.



When virus nucleocapsids are observed in the electron microscope, one often sees apparent "lumps" or clusters on the surface of the particle. These are usually protein subunits clustered around an axis of symmetry, and have been called "morphological units" or **capsomers**.

2. Sequence identity and structural similarity

Structural Diversity Reflects Functional Diversity in Globular Proteins

In a globular protein, different segments of a polypeptide chain (or multiple polypeptide chains) fold back on each other. The folding generates a compact form relative to polypeptides in a fully extended conformation. The folding also provides the structural diversity necessary for proteins to carry out a wide array of biological functions. Globular proteins include enzymes, transport proteins, motor proteins, regulatory proteins, immunoglobulins, and proteins with many other functions.

As a new millennium begins, the number of known three-dimensional protein structures is in the thousands and more than doubles every two years. This wealth of structural information is revolutionizing our understanding of protein structure, the relation of structure to function, and even the evolutionary paths by which proteins arrived at their present state, which can be glimpsed in the family resemblances that are revealed as protein databases are sifted and sorted. The sheer variety of structures can seem daunting. Yet as new protein structures become available it is becoming increasingly clear that they are manifestations of a finite set of recognizable, stable folding patterns.

Our discussion of globular protein structure begins with the principles gleaned from the earliest protein structures to be elucidated. This is followed by a detailed description of protein substructure and comparative categorization. Such discussions are possible only because of the vast amount of information available over the Internet from resources such as the Protein Data Bank (PDB; www.rcsb.org/pdb), an archive of experimentally determined three-dimensional structures of biological macromolecules.

Key Definitions and Issues in Structural Similarity Comparison

Definition of 3D Molecular Structure:

We represent the 3D molecular structure of a protein as a collection of (possibly typed) atoms or groups of atoms in some given 3D relative placement. The placement of a group of atoms is defined by the position of a reference point (e.g. the center of a particular atom in the group) and the orientation of a reference direction. When we say that the atoms or groups of atoms may be typed, we simply mean that we may choose to label each point representing an atom or group of atoms in the structure with a tag indicating what atom or group of atoms the point is representing.

Definition of a Matching Between Two Structures Two structures match if and only if we have:

1.Correspondence—There is a one to one map between elements of the structure

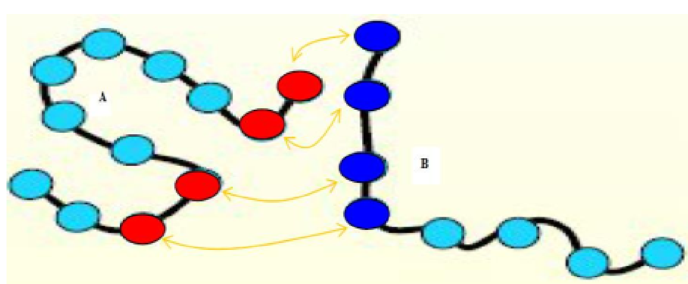
2.Alignment—There exists a rigid body transform T such that the RMSD between elements in A and those in T(B) is less than some threshold ε,

In practice a complete match of this sort between two proteins is rarely possible; in many cases of interest, two proteins may be only locally similar, may be of different sizes, or may differ structurally in other ways despite significant structurally similarity in other respects. In these cases a complete match of the two proteins is clearly too much to ask for. We can, however, hope for a partial match of two proteins.

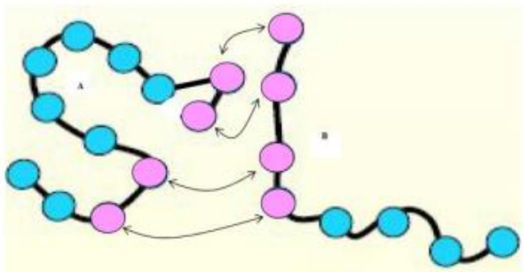
We say we have a partial matching between two proteins A and B when we have a substructure σ (A) of A and a substructure σ (B) of B such that there is a correspondence between σ (A) and σ (B) and an alignment T of A with B such that the RMSD between elements in σ (A) and those in T(σ (B)) is less than some threshold ε. We call the substructures σ (A) and σ (B) the supports of A and B, respectively. When a support is small, we refer to it as a motif. Generally, we do not require the support of a protein to be connected; the support may have two or more components which may not lie contiguously on the protein, and this can add to the challenge of the problem of finding a partial matching, as discussed below.

In formulating the problem of partial matching as above, a problem dual to that of finding the transform which minimizes RMSD arises—namely that of choosing the supports of A and B. Clearly there is a tradeoff between the allowed size of the supports of the two structures being aligned and the size of the RMSD. A common solution is to declare at the outset some maximum value ε of the RMSD and to then find the largest supports of A and B such that the RMSD between A and B with respect to those supports is less than ε.

Beyond the size of the support and the RMSD calculated from a match, there are a number of other issues that should be considered in the development of a measure of partial similarity between two proteins. For one, there may be multiple partial matches between substructures of 2 proteins. Secondly, if non-contiguous supports are permitted, one must consider the matter of whether and how to penalize to penalize for gaps in the supports of A and B, such as that depicted in the partial match below.



Third, one must consider how and whether to penalize matches where a subregion of the support of B has its orientation with respect to the backbone of B flipped relative to the orientation of the corresponding substructure of A with respect to the backbone of A, as in the picture below.



Fourth, we must decide to what extent our scoring method will adjust the score according to preference for type or backbone sequence matching. Fifth, we may wish to weight correspondences along accessible parts of the protein surface more heavily, since on average the geometry of these parts is more responsible for functional properties of the B A B A protein than the geometry of the occluded parts of the surface. Sixth, we must decide whether our similarity measure should calculate a RMSD, or arrive at its score using another similarity measure.

RMSD is by no means the only way to score similarity, and there is no consensus on what the best method is, but RMSD does have the advantage of being computationally very convenient. To offer an example of an alternative measure of similarity, below the formula for RMSD is compared with a different similarity measure used by the structure comparison software STRUCTAL. Note that RMSD is actually a dissimilarity measure (the more dissimilar the two structures being compared, the higher value it gives, so that in practice we'd want to take our measure of similarity to be 1/x, where x is the value output by RMSD). STRUCTUAL's measure, on the other hand, gives higher values when the two proteins being compared are more similar.

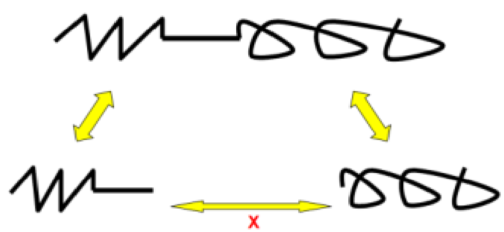
$$\min_T \sqrt{\frac{1}{|\sigma(T)|}\sum_{i\in\sigma(T)}\left(\|a_i - T(b_i)\|^2\right)}$$

RMSD dissimilarity measure

$$\max_T\left[\sum_{i\in\sigma(T)} \frac{A}{1+\left(\frac{\|a_i-T(b_i)\|}{B}\right)^2} - NGAP/2\right]$$

STRUCTAL's similarity measure

Computationally assessing protein structure similarity is a difficult problem. The difficulty can be seen as reflection of the fact that measuring partial similarity is an illposed problem; there are many ways in which two 3D structures can be similar, and depending on the application of interest, similarities between certain aspects of geometric structure may be of more interest than others. The fact that there is no single way of deciding which aspects of structure to give importance to in choosing a quantitative measure of structural similarity accounts for much of the difficulty of comparing proteins according to structure.



Whatever our choice of similarity measures, though, it is not likely to define a metric on protein structures; we cannot expect the triangle equality to be satisfied, as the picture below illustrates more clearly than words ever could.

It turns out that with respect to all partial similarity measures of interest, finding an optimal partial match between two proteins (i.e. a choice of supports, a correspondence between them, and a transformation aligning corresponding parts of the supports) is NPHard. Thus we must be satisfied with approximate/heuristic solutions to the problem. There is probably not a single best solution to computing partial matchings; rather, specific algorithms are best suited to specific applications. But even so, there are general algorithmic principles that hold across different application areas.

To close this section, we'll mention that though we often are interested in using a similarity measure more sophisticated than RMSD in computing a partial matching of two proteins, one useful method is to compute a preliminary approximate matching using RMSD and then adjust the computed transform to maximize the score of the more sophisticated similarity measure.

Applications of Structure Similarity Analysis

Though all structural similarity algorithms have a similar goal at their core, there are several different particular applications in biology today that call for somewhat different approaches.

Problem #1: Matching of Protein Structures
Given two molecules A and B, we seek substructures between A and B as large as possible while at the same time being "similar" (typically measured in RMSD).

Though the problem is stated as comparing one molecule to another, often this algorithm is used in one-to-many searches for similarity. For example, given a particular molecule, we might want to search all known proteins in the Protein Data Bank (PDB) for similarities. Or, we may want to group proteins in the PDB by doing many-to-many comparisons and clustering based on similarities.

Problem #2: Protein Classification
Besides finding similar substructures, proteins can be compared by their overall structure, i.e. classifying proteins into a hierarchy to determine similarities. Traditionally, these classifications are done manually with the aid of some automated tools, and take into account information that biologists have on the function and origin of the proteins. An example of this is the Structural Classification of Proteins (SCOP) database. It is felt that the SCOP database does a better job of classifying proteins according to structure than automated methods have

been able to thus far. However, as the number of known structures is growing rapidly, we are approaching the point where the number of new structures will be to many to be classified by hand, so good automatic methods for structure comparison and identification are becoming increasingly important.

The source of protein structures is the Protein Data Bank. The unit of classification of structure in SCOP is the protein domain. What the SCOP authors mean by "domain" is suggested by their statement that small proteins and most medium sized ones have just one domain, and by the observation that human hemoglobin, which has an $\alpha_2\beta_2$ structure, is assigned two SCOP domains, one for the α and one for the β subunit.

The shapes of domains are called "folds" in SCOP. Domains belonging to the same fold have the same major secondary structures in the same arrangement with the same topological connections. 1195 folds are given in SCOP version 1.75. Short descriptions of each fold are given. For example, the "globin-like" fold is described as *core: 6 helices; folded leaf, partly opened*. The fold to which a domain belongs is determined by inspection, rather than by software.

The levels of SCOP are as follows.

1. Class: Types of folds, e.g., beta sheets.
2. Fold: The different shapes of domains within a class.
3. Superfamily: The domains in a fold are grouped into superfamilies, which have at least a distant common ancestor.
4. Family: The domains in a superfamily are grouped into families, which have a more recent common ancestor.
5. Protein domain: The domains in families are grouped into protein domains, which are essentially the same protein.
6. Species: The domains in "protein domains" are grouped according to species.
7. Domain: part of a protein. For simple proteins, it can be the entire protein.

The folds are grouped into "classes". The classes are the top level, or "root" of the SCOP hierarchical classification. The classes are displayed something like this:

Classes:

a. All alpha proteins [46456] (284)

Domains consisting of α-helices

b. All beta proteins [48724] (174)

Domains consisting of ß-sheets

c. Alpha and beta proteins (a/b) [51349] (147)

Mainly parallel beta sheets (beta-alpha-beta units)

d. Alpha and beta proteins (a+b) [53931] (376)

Mainly antiparallel beta sheets (segregated alpha and beta regions)

e. Multi-domain proteins (alpha and beta) [56572] (66)

Folds consisting of two or more domains belonging to different classes

f. membrane and cell surface proteins and peptides [56835] (58)

Does not include proteins in the immune system

g. Small proteins [56992] (90)

Usually dominated by metal ligand, heme, and/or disulfide bridges

h. coiled-coil proteins [57942] (7)

Not a true class

i. Low resolution protein structures [58117] (26)

Peptides and fragments. Not a true class

j. Peptides [58231] (121)

peptides and fragments. Not a true class.

k. Designed proteins [58788] (44)

Experimental structures of proteins with essentially non-natural sequences. Not a true class

The number in brackets, called a "sunid", is a SCOP unique integer identifier for each node in the SCOP hierarchy. The number in parentheses indicates how many elements are in each category. For example, there are 284 folds in the "All alpha proteins" class. Each member of the hierarchy is a link to the next level of the hierarchy.

The first few folds of the 284 folds in the "All-α proteins" class are displayed something like the following.

Folds:

1. Globin-like [46457] (2)

core: 6 helices; folded leaf, partly opened

2. Long alpha-hairpin [46556] (20)

2 helices; antiparallel hairpin, left-handed twist

3. Type I dockerin domain [63445] (1)

tandem repeat of two calcium-binding loop-helix motifs, distinct from the EF-hand

Each fold is followed by a description of that fold.

The domains within a fold are further classified into superfamilies, which, in turn, are classified into families. Within a fold, domains belonging to the same superfamily are assumed to have a common ancestor. However, this ancestor is presumed to be distant, because the different members of a superfamily have low sequence identities. The two superfamilies of the "Globin-like" fold are displayed something like the following:

Superfamilies:

1. Globin-like [46458] (4)
2. alpha-helical ferredoxin [46548] (2) *contains two Fe4-S4 clusters*

No description is given for the "Globin-like" superfamily, presumably because its description is very like that of its fold, which has the same name.

Families are more closely related than superfamilies. Domains within a fold are placed in the same family if

1. they have at least a 30% similarity in sequences, or, failing that,
2. if they have some similarity in sequences, e.g., 15%, and perform the same function.

The similarity in sequence and structure is evidence that these proteins have a closer evolutionary relationship than do proteins in the same superfamily. Sequence tools, such as BLAST, are used to assist in placing domains into superfamilies and families. The four families in the "Globin-like" superfamily of the "Globin-like" fold are displayed something like the following.

Families:

1. Truncated hemoglobin [46459] (6) *lack the first helix (A)*
2. Nerve tissue mini-hemoglobin (neural globin) [74660] (1) *lack the first helix but otherwise is more similar to conventional globins than the truncated ones*
3. Globins [46463] (81) *Heme-binding protein*
4. Phycocyanin-like phycobilisome proteins [46532] (26) *oligomers of two different types of globin-like subunits containing two extra helices at the N-terminus binds a bilin chromophore*

The families in SCOP may also be referred to using a SCOP concise classification string, *sccs*, which looks like, e.g., a.1.1.2 for the "Globin" family. The letter identifies the class to which the domain belongs; the following integers identify the fold, superfamily, and family, respectively.

Within a family are protein domains. Proteins are placed in the same protein domain if they are isoforms of each other, or if they are essentially the same protein, but from different species. This is apparently done manually. The "protein domains" are further subdivided into species. ("Protein domains" are not on separate pages in the current release of SCOP; in pre-SCOP, they are on separate pages.) Here is how some of the 81 protein domains of the "Globins" family are displayed.

Protein Domains:

7. Leghemoglobin [46481]

1. Yellow lupin (Lupinus luteus) [TaxId: 3873] [46482] (17)
2. Soybean (Glycine max), isoform A [TaxId: 3847] [46483] (2)

8. Non-symbiotic plant hemoglobin [46484]

1. Rice (Oryza sativa) [TaxId: 4530] [46485] (1)

9. Hemoglobin, alpha-chain [46486]

1. Human (Homo sapiens) [TaxId: 9606] [46487] (192)
2. Human (Homo sapiens), zeta isoform [TaxId: 9606] [68937] (1)
3. Horse (Equus caballus) [TaxId: 9796] [46488] (19)
4. Deer (Odocoileus virginianus) [TaxId: 9874] [46489] (1)

The "TaxId" is the taxonomy ID number; it is also a link to the NCBI taxonomy browser, which provides more information about the species to which the protein belongs.

Clicking on a species or isoform brings up a list of domains. Here is how some of the 192 domains of the "Hemoglobin, alpha-chain from Human (Homo sapiens)" protein are displayed.

Several automated classifiers have been designed, among them are CATH (Class, Architecture, Topology, and Homologous superfamily) and FSSP (Families of Structurally Similar Proteins). As an example of how these work, the CATH protein hierarchy separates proteins at level 1 by “class” (i.e. whether the protein contains only alpha helices or beta strands or both), at level 2 by “architecture” (the gross orientation of secondary structures, currently done manually), at level 3 by “topology” (the connections between and numbers of secondary structures), and at the lowest level by “homologous superfamilies” (which takes into account structural and functional similarities between proteins).

One thing to note is that hierarchies obtained by automatic methods may be quite different from classifications designed manually because of the additional depth of knowledge that biologists have in relating proteins. Also, the splitting in the hierarchies is determined by our set of known proteins and so might by biased because there are only some proteins that we have currently been able to crystallize (i.e. determine the positions of the atoms in the molecule). Problem #3: Finding Motif in Protein Structure This problem aims to determine whether a motif, consisting of a small collection of atoms, matches anywhere in a very large protein. Note that the pieces of the motif are not necessarily connected, so we may not be able to constrain the search to consecutive atoms in the protein. Often times it is difficult to isolate such a small motif, so we augment our search by using feature “types” (i.e. requiring that we have matches between types of atoms or between types groups of atoms as well as between locations of atoms). This dramatically simplifies our problem since there are much fewer candidate sites in the protein that matches that combination of atom types.

Problem #4: Finding Pharmacaphore in Ligands

A ligand is a molecule that binds to another molecule to form a larger compound. For proteins, this can have the effect of inhibiting the proteins function or catalyzing its activities. Therefore, ligands are important in drug design. Given a set of ligands that are known to have the same activity (i.e. they all have the same effect on a protein or bind to the same site), we would like to find a substructure common to all the ligands (a pharmacaphore). Ligands are typically flexible molecules, meaning they might be in one of several conformations when they bind to the protein. Thus, for each ligand we give a set of low-energy conformations (which are more likely to react with the protein) and require that the pharmacaphore exist in at least one conformation for each ligand. This problem is one of the key problems in drug design: if we observe that a set of ligands produce the desired activity, solving this problem will hopefully elucidate the essence of the interacting substructure and allow us to design better drugs.

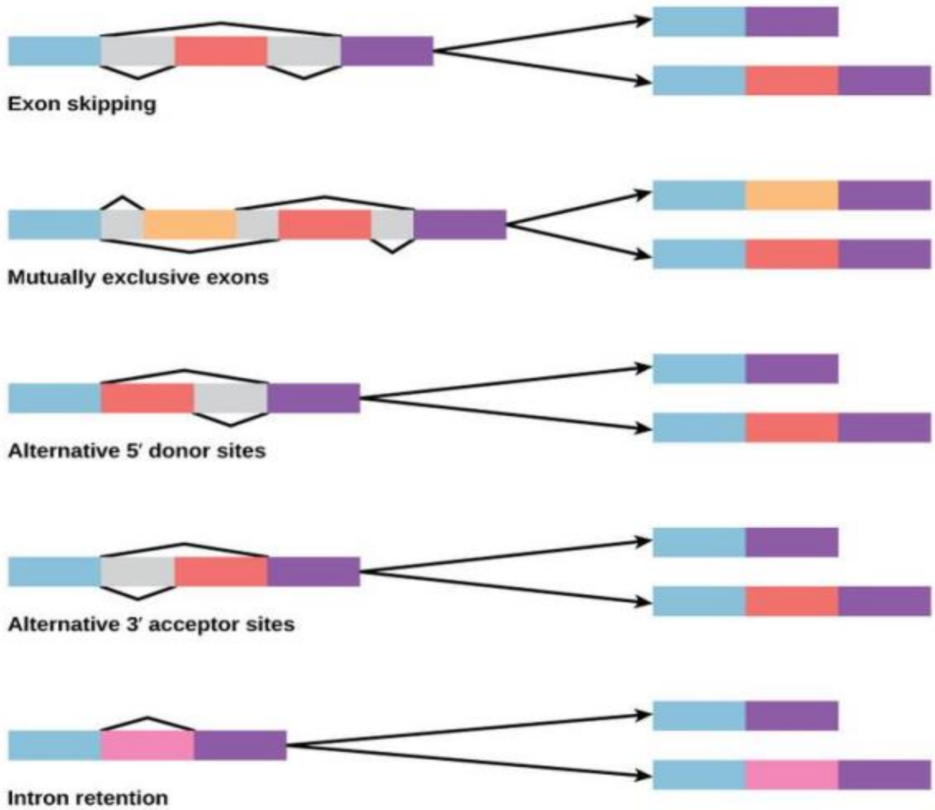
Problem #5: Search for Ligands Containing a Pharmacaphore

This problem is related to the previous, but now we are already given the pharmacaphore and would like to find all the ligands in a database that contain it. Pharmaceutical companies typically have databases of 100,000s of flexible ligands and some of their low-energy conformations. By searching for pharmacaphores with known interaction properties with a protein, we can potentially find ligands that are better. This process gives chemists a better starting point for trying to improve drugs.

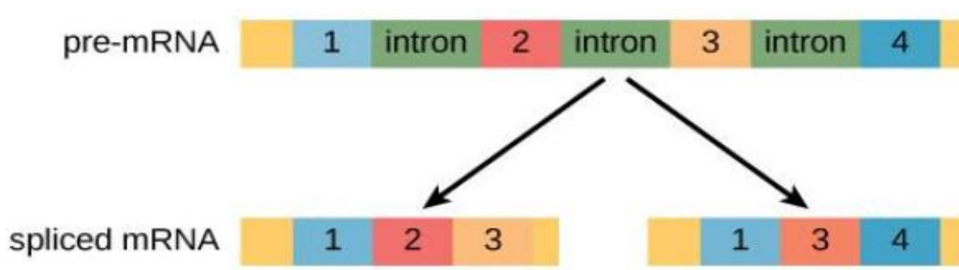
3. Protein diversity – introns, exons and inteins and exteins

RNA splicing, the first stage of post-transcriptional control

Gene expression is the process that transfers genetic information from a gene made of DNA to a functional gene product made of RNA or protein. Genetic Information flows from DNA to RNA by the process of transcription and then from RNA to protein by the process of translation. In order to ensure that the proper products are produced, gene expression is regulated at many different stages during and in between transcription and translation. In eukaryotes, the gene contains extra sequences that do not code for protein. In these organisms, transcription of DNA produces pre-mRNA. These pre-mRNA transcripts often contain regions, called **introns**, that are intervening sequences which must be removed prior translation by the process of splicing. The regions of RNA that code for protein are cal **exons** . Splicing can be regulated so that different mRNAs can contain or lack exons, i process called alternative splicing. Alternative splicing allows more than one protein to produced from a gene and is an important regulatory step in determining which functic proteins are produced from gene expression. Thus, splicing is the first stage of p transcriptional control.



Alternative Splicing: five basic modes of alternative splicing.



Pre-mRNA can be alternatively spliced to create different proteins.

Alternative Splicing

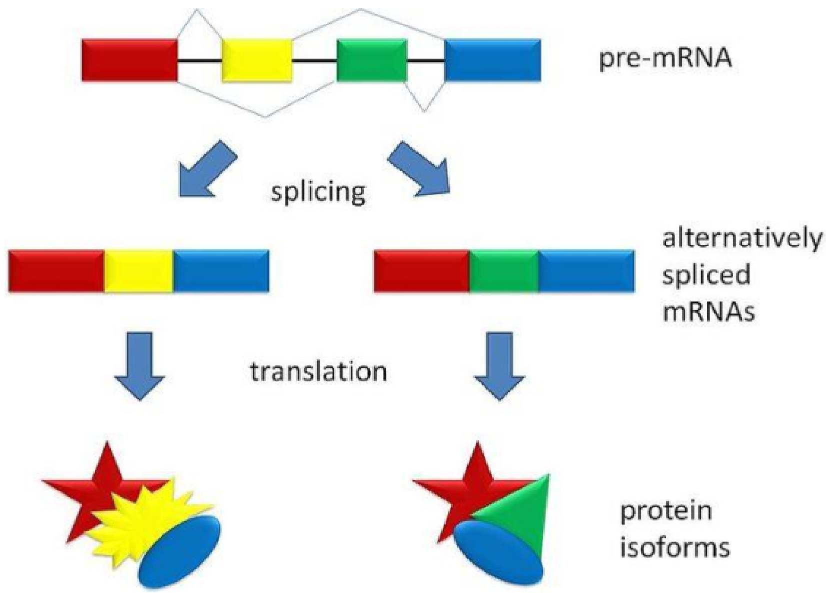
Alternative splicing is a process that occurs during gene expression and allows for the production of multiple proteins (protein isoforms) from a single gene coding. Alternative splicing can occur due to the different ways in which an exon can be excluded from or included in the messenger RNA. It can also occur if portions on an exon are excluded/included or if there is an inclusion of introns. For example, if a pre-mRNA has four exons (A, B, C, and D), these can be spliced and translated in a number of different combinations. Exons A, B, and C can be translated together or Exons A, C, and D can be translated. This results in what is called alternative splicing . The pattern of splicing and production of alternatively-spliced messenger RNA is controlled by the binding of regulatory proteins (trans-acting proteins that contain the genes) to cis-acting sites that are found on the pre-RNA. Some of these regulatory proteins include splicing activators (proteins that promote certain splicing sites) and splicing repressors (proteins that reduce the use of certain sites). Some common splicing repressors include: heterogeneous nuclear ribonucleoprotein (hnRNP) and polypyrimidine tract binding protein (PTB). Proteins that are translated from alternatively-spliced messenger RNAs differ in the sequence of their amino acids which results in altered function of the protein. This is one reason why the human genome can encode a wide diversity of proteins. Alternative splicing is a common process that occurs in eukaryotes; most of the multi-exonic genes in humans are spliced alternatively. Unfortunately, abnormal variations in splicing are also the reason why there are many genetic diseases and disorders.

Spliceosome

The splicing of messenger RNA is accomplished and catalyzed by a macro-molecule complex known as the spliceosome. The areas for ligation and cleavage are determined by the many sub-units of the spliceosome which include the branch site (A) and the 5' and 3' splice sites. Interactions between these sub-units and the small nuclear ribonucleoproteins (snRNP) found in the spliceosome create a spliceosome A complex which helps determine which introns to leave out and which exons to keep and bind together. Once the introns are cleaved and removed, the exons are joined together by a phosphodiester bond.

Regulatory Proteins

As noted above, splicing is regulated by repressor proteins and activator proteins, which are are also known as trans-acting proteins. Equally as important are the silencers and enhancers that are found on the messenger RNAs, also known as cis-acting sites. These regulatory functions work together in order to create splicing code that determines alternative splicing.



Alternative splicing can result in protein isoforms.

EXTEINS AND INTEINS

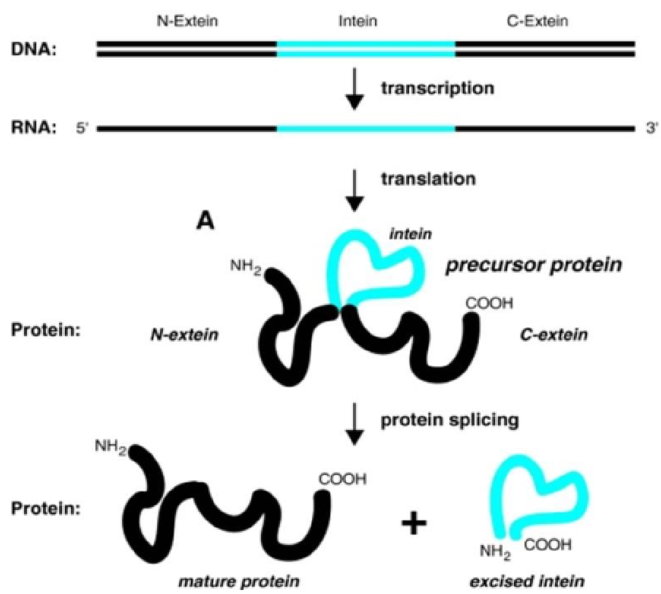
In reality, however, many proteins require additional factors to fold into their active conformation, while others remain intrinsically disordered as a requirement for their function. Furthermore, most proteins are matured, activated, inhibited, translocated, and/or degraded through the chemical modification of their side chains and backbones after protein synthesis, adding yet another layer of complexity to their structure and function.

In many cases, these post-translational modifications (PTMs) expand the chemical and structural repertoire of the canonical twenty amino acids by the addition of new functional groups to their side chains. These enzymatically applied modifications include (but are not limited to) phosphorylation, acetylation, methylation, lipidation, glycosylation, and hydroxylation, and the dynamic interplay between their addition and removal governs biological signaling.

In other cases, the primary sequence of a protein is post-translationally altered by the scission of one or more peptide bonds. This "processing" of a polypeptide chain is often carried out enzymatically by proteases and is most commonly utilized for protein degradation. However, it can also serve to activate an enzyme (e.g. the cleavage of prothrombin to yield thrombin), remove a translocation signal (e.g. signal peptide removal during antibody secretion), or mature a protein into a functional state (e.g. the conversion of proinsulin into active insulin). Remarkably, several classes of proteins can modify their own peptide backbones, and these modules are referred to as auto-processing domains. These useful auto-processing domains comprise a conserved family of proteins known as inteins.

PROTEIN SPLICING: A WIDE SPREAD POST TRANSLATIONAL MODIFICATION

An intein (*in*tervening pro*tein*) carries out a unique auto-processing event known as protein splicing in which it excises itself out from a larger precursor polypeptide through the cleavage of two peptide bonds and, in the process, ligates the flanking extein (*ex*ternal pro*tein*) sequences through the formation of a new peptide bond. This rearrangement occurs post-translationally (or possibly co-translationally), as intein genes are found embedded in frame within other protein-coding genes. Furthermore, intein-mediated protein splicing is spontaneous; it requires no external factor or energy source, only the folding of the intein domain.



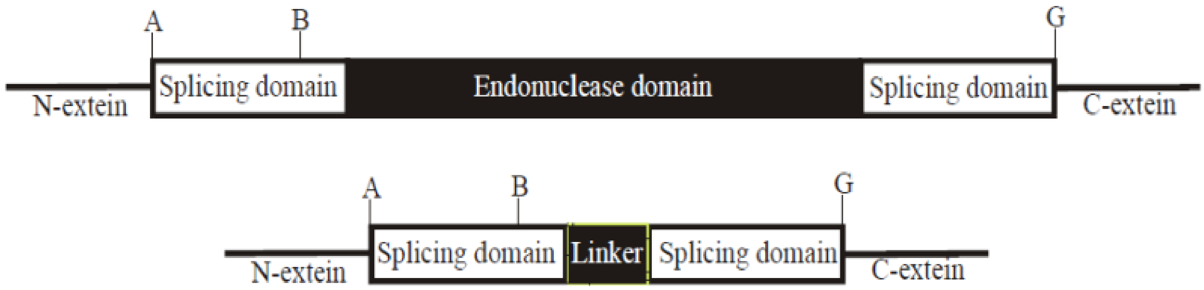
DISCOVERY

The first intein was discovered in 1988 through sequence comparison between the *Neurospora crassa*[and carrot vacuolar ATPase (without intein) and the homologous gene in yeast (with intein) that was first described as a putative calcium ion transporter.

In 1990 Hirata *et al* demonstrated that the extra sequence in the yeast gene was transcribed into mRNA and removed itself from the host protein only after translation. Since then, inteins have been found in all three domains of life (eukaryotes, bacteria, and archaea) and in viruses.

Within the database of all known inteins (Inbase), 113 known inteins are present in eukaryotes with minimum length of 138 amino acids and maximum length of 844 amino acids. The first intein was found encoded within the VMA gene of *Saccharomyces cerevisiae*. They were later found in fungi (ascomycetes, basidiomycetes, zygomycetes and chytrids) and in diverse proteins as well. A protein distantly related to known inteins containing protein, but closely related to metazoan hedgehog proteins, has been described to have the intein sequence from Glomeromycota. While in eubacteria and archea, there are 289 and 182 known inteins.

STRUCTURE



Inteins consists of two domains

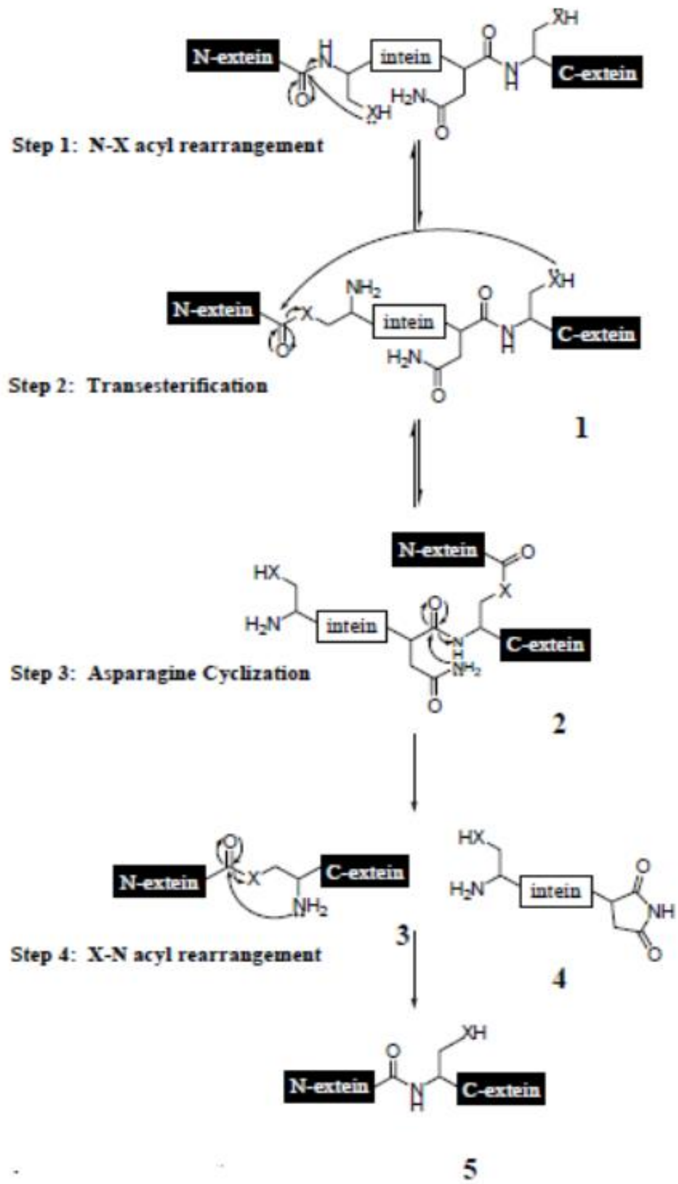
* Splicing domain which comprises the N and C terminals.

* An endonuclease domain.

Certain amino acids are highly conserved within the protein splicing domain. The first amino acid on the N-terminal intein (block A) and the C-terminal extein is either serine, cysteine, or threonine, although an alanine residue has sometimes been observed at the N-terminal splice junction.1 The amino acid on the C-terminal of the intein (block G) is most commonly asparagine, preceded by a histidine.7 Finally, a Thr-x-x-His sequence in block B of the N-terminal intein is usually observed

Inteins consisting of only the self splicing domain are called <u>mini-inteins.</u>

MECHANISM OF PROTEIN SPLICING



Step 1: Formation of an ester intermediate by an N-O acyl rearrangement of the conserved Ser /Cys residue at the upstream splice junction.

Step 2: Formation of a branched intermediate by transesterification

Step 3: Intein excision by peptide bond cleavage coupled to succinimide formation involving the conserved Asn residue at the downstream splice junction.

Step 4: Spontaneous O-N acyl rearrangement in formation of peptide bond between 2 exteins.